Applying PCA for Traffic Anomaly Detection: Problems and Solutions

Daniela Brauckhoff, Kavé Salamatian, Martin May

Applying PCA for Traffic Anomaly Detection: Problems and Solutions

Daniela Brauckhoff
ETH Zurich
Zurich, Switzerland
brauckhoff@tik.ee.ethz.ch

Kave Salamatian
Lancaster University
Lancaster, England
kave@lancaster.ac.uk

Martin May
ETH Zurich
Zurich, Switzerland
may@tik.ee.ethz.ch

***Abstract*—Spatial Principal Component Analysis (PCA) has been proposed for network-wide anomaly detection. A recent work has shown that PCA is very sensitive to calibration settings, unfortunately, the authors did not provide further explanations for this observation. In this paper, we fill this gap and provide the reasoning behind the found discrepancies.**

First, we revisit PCA for anomaly detection and evaluate its performance on our data. We develop a slightly modified version of PCA that uses only data from a single router. Instead of correlating data across different spatial measurement points, we correlate the data across different metrics. With the help of the analyzed data, we explain the pitfalls of PCA and underline our argumentation with measurement results. We show that the main problems that make PCA difficult to apply are (i) the temporal correlation in the data; (ii) the non-stationarity of the data; and (iii) the difficulty about choosing the right number of components. Moreover, we propose a solution to deal with the most dominant problem, the temporal correlation in data. We find that when we consider temporal correlation, PCA detection results are significantly improved.

I. Introduction

Principal Component Analysis (PCA) has been first proposed as a method for traffic anomaly detection in [**?**]. While being known in other domains before, Lakhina et al made its application very popular in the networking community. Subsequent publication confirmed the excellent performance of PCA and proposed extensions to it [**?**]. Only recently, it has been shown by Ringberg et al [**?**] that PCA is very sensitive to its parameter settings. The authors have reported about instability problems encountered when using PCA, however, they failed in providing precise reasons for their observation.

In this paper, we will provide the missing explanations for the encountered problems. In particular, we revisit PCA-based approaches for anomaly detection from a signal processing point of view. During the application of the PCA method on our dataset, we found similar inconsistencies as those reported in [**?**]. Further investigating the results, we found that the main problem of PCA, as used today, is that it does not consider the temporal correlation of the data. The fact that the data is temporally correlated indeed breaks the underlying hypothesis of the PCA-based anomaly detection method. The main contributions of this paper are that we (i) show what kind of problems arise when PCA is not carefully applied to anomaly detection; and (ii) provide a profound theoretical explanation for the encountered problem; and (iii) provide correction to the previously published method and alter it to a efficient anomaly detection mechanism. We validate our improved method by applying it to real network traffic with well known and identified anomalies.

The paper is structured as follows. In section II, we revisit the basics of anomaly detection methods and introduce the dataset used throughout this paper. Then, we report on the application of PCA to this data. The first results show very poor performance of the detector and by investigation the reason for it, we find the some basic features necessary for anomaly detection are not provided with PCA. For example, we find that the decision variable does not have a mean of zero - unfortunately PCA should be applied to zero mean random variables only. We propose two mechanisms that help to get the mean closer to zero and we'll show how these mechanisms are applied to the dataset.

In section III, we revisit the PCA theory and develop its basic properties. We thereafter extend the analysis to stochastic processes and explain why the simple PCA is not applicable and should be replaced by Karhunen-Loeve (KL) expansion. We describe this expansion an develop a Galerkin-based approach to calculate the KL expansion from a finite number of samples of the process. We then use the KL expansion to develop the predictive model that could be used for anomaly detection. We present the characteristics of the predictive models and propose a new methodology for choosing the optimal number of components in the KL expansion.

Section IV applies the KL extension method developed in section III to our data and examines therewith the reasons for the bad performance of classic PCA. We are able to validate that the source of the poor performance really is the temporal correlation. We further show that the non-stationarity is a critical issue and recalibration of the model is mandatory for good anomaly detection performance. Finally we describe the predictive models used for anomaly detection and open way for a signal processing approach to anomaly detection where the main challenge becomes to design filters that will be adapted to anomaly features.

II. Practical exercise: applying PCA for anomaly detection on our network

The main goal of our research is to develop efficient anomaly detection methods that are convenient for medium-

sized ISPs (as for example, the SWITCH network, AS559). One of the most frequently used method for anomaly detection is the Principal Component Analysis based subspace method developed in [**?**], [**?**]. It was therefore, an obvious choice for a starting point to apply this method to our data. In this section, we give an introduction to anomaly detection in general, describe our dataset, and present the results of the PCA-based anomaly detection method on this data.

A. About anomaly detection

An anomaly detector consists essentially of two components: (i) an entropy reduction component and (ii) a decision component applying statistical tests to a decision variable issued from the first step. The entropy reduction step is here to simplify the second step.

A predictive model that forecasts the value of the parameter to monitor as $\hat{\mathfrak{X}}[k]$ is frequently used to build the entropy reduction. This results in an error signal that has a smaller entropy than the initial signal, but still retains the information required for anomaly detection. A decision variable $D[k]$ used in the second step, is derived as a function of the prediction error $\mathbf{e}[k] = \mathfrak{X}[k] - \hat{\mathfrak{X}}[k]$. In PCA based method, we build the predictive model assuming the projection in an orthogonal subspace obtained through application of PCA is a good predictor of the signal (see section III for a detailed description).

In the second step, we apply a statistical test to a decision variable that depends on the prediction error. Generally, the statistical test step checks the null hypothesis $\{H_0$: the decision variable is compatible with a prediction in conformity with observation$\}$, *i*.e.,we reject the null hypothesis if the likelihood of the decision variable is below a threshold. The statistical test needs to have the distribution of the decision variable to obtain its likelihood. Two types of errors can occur during the statistical test: false negatives when one assumes that the null hypothesis is valid when there is an anomaly (the likelihood of observing the given variable being smaller than the threshold even if they have been an anomaly), and false alarm when deciding to refute the null hypothesis when there is no anomalies. The Neyman-Pearson theorem about statistical tests [**?**] defines a fundamental trade-off between false alarm probability and true negative probability; larger thresholds lead to lower false negative rates but larger false alarms rate and smaller thresholds result in higher false negative rates and smaller false alarms rate. The Receiver Operating Characteristics (ROC) Curve combining the two parameters in one value [**?**] captures this essential trade-off. The ROC curve is indeed a good performance metric for the two steps of an anomaly detector. A good entropy reduction technique is one that generates a decision variable with a good detection vs. false alarm rate tradeoff, *i*.e.,achieves a high detection rate with few false alarms.

For PCA based anomaly detectors, [**?**] proposed to use a non-linear function of the squared error $Q[k] = \mathbf{e}[k]^T\mathbf{e}[k]$ as decision variable $D[k]$

$$D(Q[k]) = \frac{Q[k]}{\theta}^h \tag{1}$$

[scale=0.5]plots/ROCcurve1.eps

Fig. 1. ROC curve resulting from application of PCA method to three variation of traffic metrics.

This non-linear function has been tailored to make its distribution converging as closely as possible to a gaussian distribution. In [**?**], the authors refer to the work of Jensen [**?**] showing that, under the hypothesis that elements of $\mathbf{e}[k]$ are statistically ***independent*** and follow a gaussian distribution, the distribution of $D(Q[k])$ converges to a normal distribution with known mean and variance. [**?**] gives formulas for deriving from PCA characteristics the mean and variance of $D[k]$, as well as the parameters θ and h. It is therefore, possible to normalize $D[k]$ to a gaussian random variable with zero mean and a variance of one. In the forthcoming, we will use the normalized $D[k]$ as decision variable. Because of space restrictions, we are not giving all the details of this formula in this paper, but refer the reader to [**?**], [**?**] and [**?**].

B. Data Set

We use for our experiments 3 weeks of Netflow data coming from one of the peering links of a medium-sized ISP (SWITCH, AS559). These data were recorded in August 2007 and comprised a variety of traffic anomalies happening in the daily operation such as network scans, denial of service attacks, alpha flows, *etc*. In this dataset we distinguish between incoming and outgoing traffic, as well as UDP and TCP flows. For each of these four categories, we computed seven commonly used traffic features: byte, packet, and flow counts, source and destination IP address entropy, as well as unique source and destination IP address counts. All metrics were obtained by aggregating the traffic at 15 minute intervals resulting in a 28×192 data matrix per measurement day.

We applied the anomaly detection methods to a matrix $\tilde{\mathbf{x}}$ of size $t \times f$ with t observations from f features. We used the first two days (394 time samples) of this dataset for model calibration. Thereafter, we applied the anomaly detection method to all time samples. anomalies in the data were using available manual labeling methods: visual inspection of time series and top-n queries on the flow data. This resulted in 28 detected anomalies event in UDP and 73 detected in TCP traffics.

It is noteworthy that our dataset is different in nature from the one used in [**?**], [**?**]. There, the data were collected from different network border routers. Also, [**?**] only analyzes the traffic volume. In the follow-up publication [**?**], the authors extended their dataset with the 4 entropy values of the source and destination IP addresses as well as the source and destination port numbers. In this work, we collected the data at a single link. As in the work of Lakhina *et al.*, the observed metrics are correlated in time and in space. Our observations and statements are valid whenever both spatial and temporal correlation exist; they can be extended to the case studied by [**?**], [**?**].

C. PCA application results

We then applied the methodology described in [**?**] to our data. With this method, we first apply the classical PCA as described in section III-A to the vector of metrics $\mathbf{x}[1:194]$ containing the first two days of metrics. For this purpose, we derive a spatial correlation matrix as $\hat{\Gamma} = \frac{1}{193}\mathbf{x}[1:194]\mathbf{x}[1:194]^T$. Then, we apply the SVD decomposition to the data, resulting in a basis change matrix. We construct a model using only the 8 top principal components (out of 28 possible). We choose component numbers following the methodology proposed in [**?**], [**?**] to englobe more than 95% of the variance in the initial metric (for detailed description see section III). We compute the term $Q[k] = \mathbf{e}[k]^T\mathbf{e}[k]$ from prediction errors, and we derived the normalized decision variable $D(Q[k])$ as described above.

Because of lack of space, we show only the results for the UDP traffic. Results for TCP traffic are analogous and are available in [**?**]. The ROC curve obtained by strict application of the PCA method as described in [**?**] leads to the figure labeled "Before mean removal" in Fig. 1.

Unexpectedly, the ROC curve showed very bad anomaly detection performance for the classic PCA method ("Before mean removal"). At best, we achieve a detection rate of 90% and this comes with the burden of 49% of false alarms. We detect only 15% of anomalies without any false alarm. The results obtained for TCP are even worse. Clearly, with such a performance this anomaly detection method is not useful in practice. After checking all the steps and ensuring that we closely followed the methodology proposed in [**?**], we had to back trace all causes that could have led to the bad anomaly detection performance.

On general terms, the results are in accordance with what was reported in [**?**]. It shows that tuning PCA to operate effectively in practice is difficult. Specifically, it shows high sensitivity of the false positive rate to the number of principal components chosen and relates this to the possibility that large anomalies pollute the normal subspace. Even [**?**] acknowledged this last point when it used another decision variable in place of the one defined in Eq. 1 to deal with the normal space pollution. However, [**?**], [**?**] failed to identify the causes of the problems they described. Particularly, a misleading term generated some confusion. In linear filter theory, any linear filter separates the space of signals into two orthogonal and dual subspaces: a subspace of the signals filtered by the filter and a subspace of the signal that pass through the filter. Calling the PCA method the subspace method is a kind of tautology. Any entropy reduction technique based on a predictive linear filter is a subspace method; PCA is just one subspace method with a very particular subspace structure. This means that in all methods, a component of the anomaly that maps to the normal space can pollute it. This is why the misdetection/false alarm trade)-off occurs! Stating that the PCA method has stability problems because of normal subspace pollution is a trivial statement as any failing prediction-based anomaly detector is failing because of normal subspace pollution! Analyzing if PCA is a suitable method or not requires less trivial reasoning. One need first to evaluate the assumption of the PCA method itself to ensure that the problem is not coming from bad application of the method. This is indeed the core contribution of our research.

D. Fault diagnostics of classic PCA

We began to back trace the inconsistencies by verifying if the normalized anomaly decision variable (Fig. 2) was following our theoretical expectations. We were expecting to have a decision variable with mean zero and variance 1. Actually, we observed that the mean of the decision variable over the 192 values used for calibration was 10.58 in place of zero! This bias could elucidate why the anomaly detection performance was bad as we apply the thresholds with an assumption of zero mean. Moreover, we found (see Fig. 2) that the convergence to a gaussian distribution is dubious, especially for the tails of the distribution where anomaly detection methods are looking at.

The first possible explanation was a well known argument, that was not stated in previous papers on anomaly detection: PCA should be applied to zero mean random variables. Meaning that one needs to first get the mean as close as possible to zero before applying PCA. We describe later in this paper how a mean close to zero can be obtained even for real-time operation. We verified if ignoring the non-zero-mean could have lead to our observations. We show in Fig. 1 the ROC curve obtained by applying PCA to the metrics after reducing the mean value. This second ROC curve shows a significant improvement of the false alarm rate. But still, the detection rate performance is below expectation. We checked the bias in the decision variable and found that it is reduced but is still important as the mean remains equal to 4.8 in place of zero.

A closer look at the proof of convergence in [**?**] shows that even if the convergence is robust toward non-gaussianity of the underlying variables, it is heavily dependent on the independence condition between the terms in $\mathbf{e}[k]$. We verified if we can validate this condition. We estimated the correlation between the error terms and found that there is a high correlation between some of these terms. So the hypothesis of independence should be fully rejected! We conclude that the observed bias comes therefore from the lack of independence between error terms.

However, PCA theory predicted that these errors terms should be independent. By pursuing further the investigation, it appeared that the problem is coming from the method proposed in [**?**], [**?**]. They applied PCA to the spatial correlation alone (see above the formula for obtaining $\hat{\Gamma}$) and did not remove the temporal correlation that is intrinsic in the used data. This temporal correlation leads to the prediction error terms and hence to the observed correlation.

E. Removing daily and weekly pattern

To check if we were going in the right direction, we tried to reduce the temporal correlation by removing the daily and weekly pattern. The observed metrics exhibited, as expected,

a diurnal and a weekly pattern so that they were clearly non-stationary. Fortunately, this pattern is highly predictable. We used the first week of traffic as a reference to derive a weekly profile. We computed it by applying a smoothing filter that consisted of a centered moving average windows of size 20, to the observed metrics during a single week. This resulted in a smoothed version of the traffic metrics that we used as weekly profile. We subtracted this profile from the observed metrics to remove the daily and weekly pattern. However, the resulting signal has still a non-zero mean. We explained earlier that it is essential to apply PCA on zero mean signals. Whenever removing the mean for a given dataset is a trivial task, it becomes more difficult to remove for an online signal in real time. For this purpose, we used a zero-mean filter that estimates the mean from a window of 20 past samples and subtracts it from the signal. This yields signals that have a mean very close to zero.

We applied the same PCA method as above to the metrics after removing the weekly pattern and the reduction of the mean. This resulted in the third ROC curve also plotted in Fig. 1. This ROC curve shows a major improvement in anomaly detection quality. We achieve a detection rate of above 95% with 22% of false alarm rate. Moreover, the bias in the decision variable decreased from 10.58 to 2.2. This validated the hypothesis that the problems we were seeing are coming from the temporal correlation. A verification of the independence condition on error terms showed that the correlation decreases significanlt, but it remains too high to accept the independence assumption. Similar results for TCP traffic can be found in [**?**].

As explained before, the main goal of the entropy reduction step is to generate a decision variable that could thereafter lead to suitable ROC curves. It seems that applying PCA as developed in [**?**], [**?**] fails to achieve this goal. The main reason is the inability of PCA to generate a set of independent error terms. In the following, we revisit the theoretical foundation for PCA and examine PCA from a signal processing point of view.

III. What is PCA: A signal processing view

A browse in the literature shows two closely related but different interpretations of PCA:

- As an efficient representation that transforms the data to a new coordinate system such that the projection on the first coordinate contains the greatest variance, the projection on second coordinate has the second greatest variance, and so on.
- As a modeling technique using a finite number of terms of an orthogonal serie expansion of the signal with uncorrelated coefficients.

Interestingly, the literature mainly motivates the application of PCA to networking anomaly detection by the first interpretation. However, this application is indeed following the two last interpretations. This has resulted in some erroneous interpretation and practices that have widely spread among the community. We are devoting this paper to describing these erroneous practices and to present a way of correcting them. Let's describe the above two interpretations with more details.

A. PCA : a suitable data representation

Let's suppose that we have a column vector of correlated random variables $\mathbf{X} = (X_1, \ldots, X_K)^T \in \mathbb{R}^k$. One observes these random variables through N independent realization vectors $\mathbf{x}^i = (x_1^i, \ldots, x_K^i)^T,\ i = 1, \ldots, N$ and arranges them in a $N \times K$ observation matrix $\mathbf{x}$ with each row containing an observation vector $\mathbf{x}^i$. We are searching for the most "**suitable**" non-canonical basis $(\mathbf{e}_1, \ldots, \mathbf{e}_K)$ for the vector space $\mathbb{R}^K$ to represent the random variables $\mathbf{X}$.

For the class of random variables that are linear (they can be decomposed to a linear combination of independent linear random variables) and have as sufficient statistics their means and covariances (*i*.e.,means and covariances entirely describe their joint probability distributions.), the most suitable basis is the one that maximizes the variance of each projected component. One very popular case where these two assumptions hold is when $(X_1, \ldots, X_K)$ are jointly gaussian. Nonetheless, the literature is full of examples where using such an orthonormal basis results in erroneous interpretation because the linearity or the sufficiency of mean and covariance is not valid.

Under assumption of linearity and sufficiency of mean and variance, the most suitable basis is $(\phi_1, \ldots, \phi_K)$, where ϕ_i is an eigenvector of the covariance matrix of $\mathbf{X}$ defined as $\mathbf{\Sigma} = \mathbb{E}\{(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T)\}$ (μ is a column vector containing the means of X_i). We derive these eigenvectors by solving the following linear equation:

$$\Sigma\phi_i = \lambda_i\phi_i \tag{2}$$

where λ_i are the eigenvalues of the covariance matrix. As the covariance matrix is positive definite, this equation has at most K positive eigenvalues and K different orthonormal eigenvectors. The basis change matrix $U = [\phi_1, \ldots, \phi_K]$ contains in its columns, the eigenvectors ϕ_i. Solving the above problem is called in matrix theory the Singular Value Decomposition (SVD) of the covariance matrix.

It is noteworthy that U is a basis change matrix only when $\mathbf{X}$ is zero mean, and in general one has to work with $\tilde{\mathbf{X}} = \mathbf{X} - \mu$ in place of $\mathbf{X}$, *i*.e.,the coordinate change is $\tilde{\mathbf{y}} = U\tilde{\mathbf{x}}$. This last point is frequently overlooked in the literature, and not taking care of it could lead to large errors when using PCA[1]. In the forthcoming we will assume that we have taken care of this obvious precaution so we can drop the $\sim$. For real time operation, removing the mean can be approximated by using a zero-mean filter as described in section II-E.

After applying PCA one can rewrite the initial vector of random variables $\mathbf{X}$ in the new coordinate system as:

$$\mathbf{X} = \sum_{i=1}^{K} Y_i\phi_i \tag{3}$$

[1]It is noteworthy that even if [**?**], [**?**] did not state clearly the necessity of removing the mean, they have used zero mean signals in their implementation code

where Y_i are jointly independent random variables with $\mathbb{E}\{Y_i\} = 0$ and $\mathbb{V}\text{ar}\{Y_i\} = \lambda_i$. PCA replaces the correlated random variables $\mathbf{X}$ by a vector of independent random variables $\mathbf{Y}$ that are linearly equivalent. The independence of Y_i is therefore an ***essential*** property as this is the main reason that PCA representation is "suitable".

The above discussion remains theoretical, and in practice one has a set of observations and has to pick the suitable basis. Whenever the dataset under study is not flagrantly in contradiction with the conditions of mean and variance sufficiency, and linearity we can apply PCA and find a convenient representation of the data.

First one has to estimate the covariance matrix using the popular sum of product formula $\hat{\Sigma} = \frac{1}{N-1}\mathbf{x}^T\mathbf{x}$. Because of independence between observations this gives a reliable estimation. Thereafter applying the SVD factorization is just a straightforward and mechanical step that could provide the needed basis as well as the basis transform matrix.

B. Extension of PCA to a vector of stochastic processes

Before describing PCA as a modeling technique, let's extend it to stochastic processes. This extension is mandatory as the signals used for anomaly detection are samples of stochastic processes that have temporal as well as spatial correlations. Let's assume we have a K-vector of zero mean stationary stochastic processes $\mathbf{X}(t) = (X_1(t), \ldots, X_K(t))^T$ with a covariance functions $\sigma_{i,j}(\tau) = \mathbb{E}\{X_i(t)X_j(t-\tau)\}$ defined over an interval $[a, b]$. The extension [2] to multi-dimension of the Karhunen-Loeve (KLT) theorem [**?**] states that one can rewrite these processes as a serie expansion (named the KL expansion):

$$X_l(t) = \sum_{i=1}^{K}\sum_{j=1}^{\infty} Y_{i,j}^l \Phi_{i,j}(t) \tag{4}$$

where $Y_{i,j}^l$ are pairwise independent random variables and $\Phi_{i,j}(t)$ are pairwise orthogonal **deterministic** (non-random) functions defined on $[a, b]$, *i*.e.,$\int_a^b \Phi_{i,j}(t)\Phi_{m,n}^*(t)dt = 0$ for $i \neq m$ or $j \neq n$. Generally, the basis function $\Phi_{i,j}(t)$ are rescaled such that $\int_a^b |\Phi_{i,j}|^2(s)ds = 1$.

This theorem extends the PCA to a vector of stochastic process. Eq. 4 is the equivalent of Eq. 3. The family of deterministic functions $\Phi_{i,j}(t)$ is an orthonormal basis for the space of linear stochastic processes and the random variables $Y_{i,j}^l$ are coordinates of the stochastic process $X_l(t)$ in this new space. We can formally derive the basis functions $\Phi_{i,j}(t)$ by solving the following set of linear integral equations:

$$\sum_{i=1}^{K}\int_a^b \sigma_{i,l}(s)\Phi_{i,j}(s-t)ds = \lambda_{l,j}\Phi_{l,j}(t),\ j > 0.$$

These complex equations are simply the equivalent of equations 2. The random variables $Y_{i,j}^l$ are obtained by projecting each stochastic process over an eigenfunction:

$$Y_{i,j}^l = \int_a^b X_l(s)\Phi_{i,j}(s)ds$$

The KL expansion considers the temporal correlation (between time t and $t+\tau$) as well as the spatial correlation (between process $X_i(.)$ and $X_j(.)$). This results in a more complex analysis than the simple PCA described earlier. However, this higher complexity is unavoidable because of the temporal correlation. Not taking it into account could lead to the errors described in section II-C.

In practice, we have only access to a finite set of samples observed each T time unit from the vector of stochastic processes and we have to implement the KL expansion only using them. In the forthcoming, we will use the notation $[k]$ to represent the discrete version of a time continuous process sampled at time kT. Let's assume that we have n samples of the multidimensional stochastic process and the covariance values $\sigma_{i,j}(\tau)$ can be assumed as negligible for $\tau > NT$. We can therefore truncate the KL expansion to N terms in place of the infinite number of terms needed normally. The Galerkin method [**?**] transforms the above integral equations to a matrix problem that could be solved by applying the SVD technique. This makes possible the derivation of KL expansion using only a finite number of samples. The Galerkin method generates a set of eigenvectors in a KN dimensional vector space, that are time-sampled version $\Phi_{i,j}[k] = \Phi_{i,j}(kT)$ of the originally continuous function $\Phi_{i,j}(t)$. Finally, we obtain a discrete version of the KL expansion as :

$$X_l[k] = \sum_{i=1}^{K}\sum_{j=1}^{N} Y_{i,j}^l \Phi_{i,j} j[k]. \tag{5}$$

We first have to estimate the spatio-temporal correlation matrix. Let's construct a $KN \times (n-N)$ observation matrix:

$$\mathbf{x} = \begin{pmatrix} x_1(1) & \ldots & x_1(n-N) \\ x_1(2) & \ldots & x_1(n-N+1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \ldots & x_1(n) \\ x_2(1) & \ldots & x_2(n-N) \\ \vdots & \ddots & \vdots \\ x_2(N) & \ldots & x_2(n) \\ \vdots & \ddots & \vdots \\ x_K(1) & \ldots & x_K(n-N) \\ \vdots & \ddots & \vdots \\ x_K(N) & \ldots & x_k(n) \end{pmatrix}$$

The matrix $\hat{\Sigma} = \frac{1}{n-N-1}\mathbf{x}^T\mathbf{x}$ is a $KN \times KN$ matrix that contains all the needed spatio-temporal covariance estimates. It is noteworthy that because of temporal correlation one needs more data to estimate correctly the covariance here than for the indepedent case we had in section III-A.

The Galerkin method consists of applying PCA as described in section III-A to this large matrix. This results in KN eigenvectors of length KN $\Phi_{i,j}[.]$ that are used to construct a basis transform matrix U. The coefficients $Y_{i,j}^l$ are obtained by applying the basis change transform $\mathbf{y} = U\mathbf{x}$. As can be seen applying KL expansion to K stochastic process entails

[2]The KL theorem was initially defined for one dimensional stochastic processes

diagonalizing a $KN \times KN$ matrix (in place of a $K \times K$ matrix in section III-A). However, this added complexity is unavoidable when one has to deal with correlated observations.

C. PCA as a modeling method

Up to now, we described the KL expansion as a tool for creating an equivalent (in probability) and suitable representation of vector of stochastic processes. If we neglect some of the smaller terms of the expansion (terms with small values of $\mathbb{V}\text{ar}\{Y_{i,j}^l\}$), we obtain a linear approximation of the initial process in a smaller dimension vector space. The discrete expansion in Eq. 5 is therefore approximated as:

$$\hat{X}_l(kT) = \sum_{i=1}^{L}\sum_{j=1}^{M} Y_{i,j}^l \Phi_{i,j}^k. \tag{6}$$

where $M < N$ and $L < K$. This approximation has a noteworthy optimality property. Among all approximations defined over a linear space of dimension LM, this is the linear approximation with the smallest approximation error variance ($\mathbb{V}\text{ar}\{X(t) - \hat{X}(t)\}$. The basis change transform becomes a $KN \times LM$ matrix U_{LM} that contains the LM eigenfunctions $\Phi_{i,j}[.]$ in its columns. This is the theoretical basis to use the KL expansion as a non-parametric and generic technique for modeling a large class of processes where we cannot reject the linearity and sufficiency of mean and variance (see section III-A).

The non-parametric nature of the above modeling technique is simultaneously its strength and Achilles heel; a non-parametric method is not based on any precise form of the distribution (out of the linearity and the mean and variance sufficiency) meaning it is more robust. At the same time being non-parametric means that no prior knowledge can be incorporated into the model.

Before going further, let's first give some details about the obtained model. The expansion in Eq. 6 provides a synthesis method for generating an approximated process $\hat{X}_l[k]$ by a bank of ML filters with Finite Impulse Response equal to $\Phi_{i,j}[k], k = 0; \ldots, KN$; each filter being excited by the random variable input $Y_{i,j}^l$. By predicting the values of the realization of the KN random variables $Y_{i,j}^l$ by applying the basis change matrix to observation $\mathbf{X}[.]$, we can use this synthesis filter as a predictive filter. This is the approach followed in PCA and KL expansion based anomaly detectors. The anomaly detection literature has foreseen this approach.

Let's assume that the K linear stochastic processes in vector $\mathbf{X}[k]$ are linear processes, *i*.e.,one can represent them using a dynamic state space representation as $\mathfrak{X}[k+1] = A\mathfrak{X}[k] + \epsilon[k]$, where $\mathfrak{X}[k]$ is a KN dimension vector constructed by concatenating N vectors $(\mathbf{X}[k], \ldots, \mathbf{X}[k-N])$ and $\epsilon[k]$ is a vector of KN independent and identically distributed (iid) random variables. Now let's assume that the process vector $\mathbf{X}[k]$ is approximated by a finite KL expansion with LM terms. There is therefore a U_{LM} basis transform matrix that maps $\mathfrak{X}[k]$ into the new coordinate : $\mathfrak{Y}_R[k] = U_{LM}\hat{\mathfrak{X}}[k]$ ($\mathfrak{Y}_R[k]$ being the reduced coordinate vector of dimension LM). The inverse projection can be found through $\hat{\mathfrak{X}}[k] = U_{ML}^T \mathfrak{Y}_R[k]$ (as U^T is a Hermitian matrix). By replacing this term in the dynamic state space representation and by algebraic manipulation we obtain the following state space model for the approximated processes:

$$\begin{cases} \mathfrak{Y}_R[k+1] = A_{\mathcal{M}}\mathfrak{Y}_R[k] + B_{\mathcal{M}}\epsilon[k] \\ \hat{\mathfrak{X}}[k] = U_{ML}\mathfrak{Y}_R[k] + D_{\mathcal{M}}\epsilon[k] \end{cases} \tag{7}$$

where $A_{\mathcal{M}} = U_{LM}^T A U_{LM}$; $B_{\mathcal{M}} = U_{LM}^T[(AS + I) \mid -S]$ with $S = -A^{-1} + U_{LM}A_{\mathcal{M}}^{-1}U_{LM}^T$; $D_{\mathcal{M}} = S$. The resulting model has LM uncorrelated state variables (in $\mathfrak{Y}$) in place of KN state variables in the initial process ($\mathfrak{X}$).

The model in Eq. 7 shows precisely the effect of removing some terms from the KL expansion: this replaces the initial system dynamic by a new system with the above parameters. We can easily compare the initial system dynamic to the obtained model through three equivalent comparison approaches: impulse response, zeros and poles location, and frequency response. We can compute the poles and zeros locations by deriving the Z-transform of the transfer function of the processes. The initial process transfer function is $\frac{\mathcal{X}(z)}{\mathcal{E}(z)} = (zI - A)^{-1}$; the NK poles of the initial system are the roots of the equation $\det(zI - A) = 0$ and all the zeros are at $z = 0$. The approximated system transfer function is $\frac{\hat{\mathcal{X}}(z)}{\mathcal{E}(z)} = U_{ML}(zI - A_{\mathcal{M}})^{-1}B_{\mathcal{M}} + D_{\mathcal{M}}$;the poles being the ML roots of $\det(zI - A_{\mathcal{M}}) = 0$ and the zeros the roots of the equation $\det(B_{\mathcal{M}} + (zI - A_{\mathcal{M}})U_{ML}^T D_{\mathcal{M}}) = 0$. Having the poles and zero locations we have a full characterization of the initial and modeled system.

In practice, we frequently do not have the state-space model of the observed process to be able to apply equation 7. One can infer the state space model following the Maximum Likelihood method used in [**?**] and thereafter use the above described approach. Another simpler approach consists of using spectral estimation techniques [**?**] to derive the spectra from the initial process. This results in an estimated spectra for the initial process. We derive the approximated process $\hat{\mathfrak{X}}[k]$ easily as $\hat{\mathfrak{X}}[k] = U_{ML}^T U_{ML}\mathfrak{X}[k]$. By applying the same spectral estimation to the approximated signal one can compute the spectra for the approximation and compare it with the spectra of the initial process. We will illustrate this approach later in this paper.

This results in a criterion to select the number of terms to keep in the KL expansion. [**?**], [**?**] it suggested to choose enough terms to capture 95% of the variance of the initial process. However, this approach does not take care of the features that will remain in the approximate model. A complete criterion retains a large enough number of terms so that the approximated process exhibits essential features that are of interest for anomaly detection. We will illustrate this approach later in section IV-D. Another outcome of the model description is that it enables the model designer to use its *a priori* knowledge to introduce or emphasize features that he believes would be useful for anomaly detection. This enables a refinement approach where in a first attempt we derive a

Fig. 2. Normal plot of normalized decision variable $D[k]$ for UDP traffic obtaining over the data used for calibrating the PCA model

Fig. 3. ROC curves resulting from KL expansion to the dataset with only mean removal for different value of the temporal correlation range.

Fig. 4. Normal plot of normalized decision variable $D[k]$ for $N = 3$

model through the non-parametric KL expansion method and refine it through inspection of its features. We can refine the model could by classical control theory techniques by adding/removing/moving some of the poles or zeros such that the frequency response or the impulse response has suitable properties.

IV. Validation

In this section, we will present the results of applying the KL expansion to the dataset described in section II-B. We first confirm that the temporal correlation is indeed the source of the bad performance of PCA based anomaly detection on our dataset. For this purpose, we have derived the KL expansion of our data set after preprocessing it only by removing the mean of the metrics.

A. Validation of KL expansion

Fig. 3 shows the ROC curves obtained for different values of the temporal correlation range N. The figure is plotted in semilog to present a better comparison between different values of N. All ROC curves are obtained with enough terms to capture 95% of the total variance in the model. In all cases, no more than two expansion terms were neede for this setting. The comparison of ROC curves (see Fig. 3) shows a considerable improvement in performance of the anomaly detection with use of KL expansion with $N = 2, 3$ and thereafter a decrease for $N = 4$. The same is observed for $N > 4$. Particularly, for $N = 3$ on can detect 97% of anomalies are with no more than 16% of false alarm rate.

To verify if the problems described in section II-C are really coming from the bias in the decision variable, we checked the value of this bias. We observed that augmenting N decreases the bias from 4.8 when $N = 1$ to 0.4 when $N = 3$. We plotted in Fig. IV-A, the distribution and a normal plot of the normalized decision variable for the model with $N = 3$. With increasing N the bias decreases (even if the marginal gain saturates). Moreover, the correlation between the terms in $\mathbf{e}[k]$ decreases significantly with augmenting N. That validates our hypothesis that relates the bad anomaly detection performance of PCA to temporal correlation.

B. Effect of non-stationarity

It remains another issue to solve: for $N \geq 4$ the performance decreases even if the bias in the decision variable decreases. We investigated this observation and found that most of the mis-detections for $N \geq 4$ happen at the end of the second week and during the third week of our dataset while they were spread in the three weeks for $N < 3$. A possible explanation is the stationarity issue: when N increases, the model contains more parameters and becomes more sensitive to the stationarity of the traffic metrics. This means that we can expect that anomaly detection performance decreases with time and this decrease is more pronounced for larger N. To further check this explanation, we plotted in Fig. 5 the ROC curves obtained only on the first week of anomaly data for increasing values of N. These ROC curves are now following the theoretical intuition: with augmenting N the anomaly detection performance ameliorates. But, it still begins to decrease for $N = 10$. This is compatible with our previous hypothesis about the effect of non-stationarity on models with increasing N.

Non-stationarity and model recalibration are indeed important issues that need careful evaluation and analysis. It is out of the scope of this paper and will be the subject of a forthcoming paper.

C. Interest of weekly trend removal

We introduced in section II-E a method to remove the daily and weekly trend. We showed also in Fig. 1 that this approach has promising performance. By further applying the KL expansion to the metrics pre-processed with mean and weekly trend removal we obtain the ROC curves plotted in Fig. 6. For the sake of comparison, we added in the plot the best performance ROC curve derived before trend removal with $N = 3$. The ROC curves in Fig. 6 show that weekly trend removal alone attains a performance close to the best model obtained without weekly trend removal. One can achieve better performance by adding a KL expansion: one can detect all anomalies with only 15% of false alarms. We can detect 93% of them with less than 7% of false alarms. Moreover, one can detect 55% of anomalies without any false alarms. Analogously to what was described in section IV-B, by augmenting N to 3 the performance degrades. We investigated this and found results analogous to what was described earlier.

D. Analysis of the predictive models

We developed in section III-C an approach for analyzing the model resulting from KL expansion. Particularly, we described an approach to evaluate the model resulting from PCA or KL expansion based on inferring the signal spectra at the output of the predictive model. We will illustrate the approach on our data in this section. We are showing in Fig. 7 the spectra of our metrics estimated using the Yule-Walker approach [**?**]. We also provide the spectra of the same metric predicted by the model resulting from the KL expansion with different numbers of kept components. The figure shows that applying the popular heuristic that consists in maintaining enough components to englobe 95% of variance is not satisfactory as it results in

[scale=0.48]plots/ROCcurve4.eps

Fig. 5. ROC curves obtained over the first week of data for different values of N.

[scale=0.48]plots/ROCcurve3.eps

Fig. 6. ROC curve resulting from KL expansion on the metrics after weekly trend removal.

a spectra that is far from the initial spectra. Specifically the predicted signal contains a higher level of high frequencies. Moreover, the low frequency part of the spectra is not well modeled. This means that using the 5 component model will result in overestimating high frequencies and missing the anomalies that are happening in lower frequencies. However, using a 12 component filter gives a better approximation of the initial signal. This leads to a finer methodology for choosing how many components of the KL expansion to use. We should choose enough components to have a good approximation of the initial signal spectra particularly in the region of the spectrum where expected anomalies will occurs.

The pole-zero diagram of the initial sytem along with the a predictive model with 8 components are plotted in Fig. IV-D. The plot shows how the predictive model approximates the initial system by positioning its 8 poles and zeros in the z-plane. If an anomaly detector designer wishes to detect a particular type of anomaly signals he might use the above Pole-zeros diagram and move slightly some of the poles or zeros to emphasize specific parts of the transfer function spectra to achieve the needed goal.

The approach here becomes similar to the design of an equalizer in sound processing: the predictive filter acts as an equalizer that should filter out essential features of the anomaly signal such that these features appears in the error signal. One can then use the predictive model resulting from the KL expansion and ameliorate it to be more sensitive to specific anomalies that will happen in specific regions of the spectra. This novel anomaly detector design methodology opens new perspectives for introducing *a priori* knowledge in the design process. We will follow up this approach in our forthcoming papers.

V. Conclusion

This paper began with a very practical problem: how to apply the popular PCA method in real world anomaly detection. We found that direct application of the PCA method results in poor performance in terms of ROC curves; we investigated the problem and found that the main source of the problem is the bias coming from correlation in prediction error terms. After a detailed theoretical analysis, it appears that the correct framework is not the classical PCA but rather the Karhunen-Loeve expansion. We have presented the KL expansion and have provided a Galerkin method for developing a predictive model. This method has thereafter been applied to data traces from the Switch network and we have shown that an important improvement is attained when temporal correlation is considered. We also have developed a methodology for designing anomaly detection predictive filters and have shown its application to real data.

It is noteworthy that this paper is not claiming that PCA is the only way to do anomaly detection. By showing that PCA is based on a predictive filter - just as the Kalman filter is - we open way for a signal processing approach to anomaly detection where the main challenge becomes to design filters that will be adapted to anomaly features. To the best of our knowledge, this is the first paper to introduce such an approach and to remove completely the curtain of black magic that draped the application of PCA to anomaly detection.

Acknowledgments

The authors would like to thank Mark Crovella for his helpful discussions that have led to some aspects of this paper. We also thank Anukool Lakhina for giving access to his code and the Abilene data he used in his papers.